Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 5381–5383, June 1992
Chemistry

Encoded combinatorial chemistry

(chemical repertoire/encoded libraries/commaless code)

SYDNEY BRENNER AND RICHARD A. LERNER

Departments of Chemistry and Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines, La Jolla, CA 92037

Contributed by Sydney Brenner, March 3, 1992

ABSTRACT The diversity of chemical synthesis and the power of genetics are linked to provide a powerful, versatile method for drug screening. A process of alternating parallel combinatorial synthesis is used to encode individual members of a large library of chemicals with unique nucleotide sequences. After the chemical entity is bound to a target, the genetic tag can be amplified by replication and utilized for enrichment of the bound molecules by serial hybridization to a subset of the library. The nature of the chemical structure bound to the receptor is decoded by sequencing the nucleotide tag.

There is an increasing need to find new molecules that can effectively modulate a wide range of biological processes, for applications in medicine and agriculture. A standard way to search for novel chemicals is to screen collections of natural materials, such as fermentation broths, plant extracts, or libraries of synthesized molecules. Assays can range in complexity from simple binding reactions to elaborate physiological preparations. The screens often only provide leads, which then require further improvement either by empirical methods or by chemical design. The process is time-consuming and costly but is unlikely to be replaced totally by rational methods even when they are based on detailed knowledge of the three-dimensional structure of the target molecules. Thus, what we might call "irrational drug design"—the process of selecting the correct molecules from large ensembles or repertoires—requires continual improvement both in the generation of repertoires and in the methods of selection.

Recently there have been several developments in using peptides or nucleotides to provide libraries of compounds for discovery of leads. The methods were originally developed to speed up the determination of epitopes recognized by monoclonal antibodies. For example, the standard serial process of stepwise search of synthetic peptides has been replaced by a variety of highly sophisticated methods in which large arrays of peptides are synthesized in parallel and screened with acceptor molecules labeled with fluorescent or other reporter groups (1, 2). The sequence of any effective peptide can be decoded from its address in the array. In another approach, combinatorial libraries of peptides are synthesized on resin beads such that each resin bead contains about 20 pmol of the same peptide (3). The beads are exposed to labeled acceptor molecules. Those with bound acceptor are identified by visual inspection and physically removed, and the peptide is sequenced directly. In principle, this method could be used with other chemical entities, provided one has a sensitive method for sequence determination.

A different method of solving the problem of identification in a combinatorial peptide library is used by Houghten *et al.* (4). For hexapeptides of the 20 natural amino acids, separate libraries are synthesized, each with the first two amino acids fixed and the remaining four positions occupied by all possible combinations. An assay, based on competition for binding or some other activity, is then used to find the library with an active peptide. On the basis of this result, 20 new libraries are synthesized and assayed to determine the effective amino acid in the third position. The process is reiterated in this fashion until the active hexapeptide is defined. This is analogous to the method used in searching a dictionary: the peptide is decoded by using a series of sieves, and this makes the search logarithmic. A powerful biological method has recently been described in which the library of peptides is presented on the surface of a bacteriophage such that each phage displays a particular peptide and contains within its genome the corresponding DNA sequence (5, 6). The library is prepared by synthesizing a repertoire of random oligonucleotides to generate all combinations, followed by their insertion into a phage vector. Each of the sequences is cloned in one phage and the relevant peptide can be selected by finding those that bind to the particular target. The phages recovered in this way can be amplified and the selection repeated. The sequence of the peptide is decoded by sequencing the DNA. Another "genetic" method has been applied by Tuerk and Gold (7) and Ellington and Szostak (8), using libraries of synthetic oligonucleotides that themselves are selected for binding to an acceptor and then amplified by the polymerase chain reaction (PCR). In this case, however, the repertoire is limited to nucleotides or nucleotide analogues that preserve specific Watson–Crick pairing and can be copied by a polymerase.

The main advantages of the genetic methods reside in the capacity for cloning and amplification of DNA sequences, which allows enrichment by serial selection and provides a facile method for decoding the structure of active molecules. However, the genetic repertoires are restricted to nucleotides and peptides composed of natural amino acids, whereas a more extensive chemical repertoire is required to populate the entire universe of binding sites. In contrast, chemical methods can provide limitless repertoires, but they lack the capacity for serial enrichment and there are difficulties in discovering the structures of selected active molecules. We have now devised a way of combining the virtues of both methods through the construction of *encoded combinatorial chemical libraries*, in which each chemical sequence is labeled by an appended "genetic" tag, itself constructed by chemical synthesis. In effect, we implement a "retrogenetic" way of specifying each chemical structure.

In outline, we perform two alternating parallel combinatorial syntheses so that the genetic tag is chemically linked to the chemical structure being synthesized. In each case, addition of a monomeric chemical unit to a polymeric structure is followed by addition of an oligonucleotide sequence which is defined as "encoding" that chemical unit. The library is built up by the repetition of this process after pooling and division. Active molecules are selected by binding to a receptor, and amplified copies of their retrogenetic tags are obtained by the PCR. DNA strands with the appropriate polarity can then be used to enrich for a subset of

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

library by hybridization with the matching tags, and the process can then be repeated on this subset. Thus serial enrichment is achieved by a process of purification, exploiting linkage to a nucleotide sequence that can be amplified. Finally, the structures of the chemical entities are decoded by cloning and sequencing the products of PCR.

Design of the Code and the Genetic Tag

It is essential to choose a coding representation in such a way that no significant part of the sequence can occur by chance in some other unrelated combination. Suppose we allocate a triplet to each of the chemical units used. Then, because the method allows us to cover all combinations and permutations of an alphabet of chemical units, unless we are careful, we could find that two different combinations have closely related sequences which differ only by a frame shift and which could not be easily distinguished by hybridization. This, potentially the greatest source of errors, can be eliminated by choosing a commaless code (9). The particular commaless triplet code that we have chosen allows 20 unique representations, as shown in Table 1.

The sequences for the PCR primers must be chosen so that they do not occur within any coding segment and so that they can be readily removed from the final PCR product because we do not want them to dominate the selective hybridization. This can be achieved by building in sites for restriction enzymes with the appropriate polarity of cutting. One of the restriction enzymes should cut at a site that permits the incorporation of a biotinylated nucleotide, such as biotinyl-dUTP, into the strand complementary to the coding strand.

All of the above conditions have been met in the following design:

5′-AGCTACTTCCCAAGG[coding sequence]GGGCCCTATTCTTAG-3′
3′-TCGATGAAGGGTTCC[anticoding strand]CCCGGGATAAGAATC-5′
Sty I *Apa* I

After cleavage with both restriction enzymes we have

5′-AGCTACTTCC CAAGG[coding sequence]GGGCC CTATTCTTAG-3′
3′-TCGATGAAGGGTTC C[anticoding strand]C CCGGGATAAGAATC-5′

The internal fragment can be cloned in an appropriate vector to sequence the individuals. The terminal overhang of the *Sty* I site can be filled in with dCTP and biotinyl-dUTP (BTP) which, because an asymmetric site was chosen, will append the biotinylated nucleotides to only one of the cleavage products.

5′-AGCTACTTCCC CAAGG[coding sequence]GGGCC CTATTCTTAG-3′
3′-TCGATGAAGGGTTC BBCC[anticoding strand]C CCGGGATAAGAATC-5′

The biotinylated fragment can be bound to avidin and, after denaturation, provides the strand suitable for hybridization and selection of the appropriate coding strands:

Avidin-BBCC[anticoding strand]C

The two PCR primers are the two sequences 5′-AGCTACTTCCCAAGG (*Sty* I primer) and 5′-CTAAGAATAGGGCCC (*Apa* I primer). Adding a biotin to the 5′ end of the *Apa* I primer would allow the isolation of the whole strand containing the anticoding sequence.

We should have at least 15 nucleotides in the coding region for effective hybridization. Thus, in a library of degree $d \geq 5$, that is, composed of five or more successive chemical units, we could code each unit by a triplet. That would allow an alphabet (A) of up to 20 different units, each corresponding to one of the triplets defined above. The complexity of the combinatorial library is A^d. Libraries with a smaller degree, say $d = 3$, should be coded by sextuplets, which, in the simplest case, could be a repeated triplet. (This size is chosen because any combination of triplets still obeys the commaless condition). In the same way, the size of the alphabet can be extended by using combinations of triplets to code for the chemical units.

Table 1. Commaless code used in this study

ttt	tct	tat	tgt
TTC	tcc	tac	tgc
TTA	tca	taa	tga
TTG	tcg	tag	tgg
ctt	cct	cat	cgt
CTC	ccc	cac	cgc
CTA	cca	caa	cga
CTG	ccg	cag	cgg
att	act	aat	agt
ATC	ACC	aac	agc
ATA	ACA	aaa	aga
ATG	ACG	aag	agg
gtt	gct	gat	ggt
GTC	GCC	gac	ggc
GTA	GCA	GAA	gga
GTG	GCG	GAG	ggg

"Sense triplets" are XYZ; nonsense triplets are xyz.

A Formal Example

As an illustration we discuss how a library of degree $d = 3$ is made with an alphabet of two amino acids, glycine and methionine. In this case, we use sextuplets to give us a reasonable length of coding sequence. To make the sequences as different as possible we code each amino acid by a combination of two different triplets as follows:

Gly = CACATG, Met = ACGGTA.

Step 1. We begin with some appropriate linker, LINK, attached to some solid-state surface and synthesize the first PCR oligonucleotide sequence on one end, in the usual 3′-to-5′ direction, to give

GGGCCCTATTCTTAG–LINK

Step 2. This product is divided into two aliquots for parallel synthesis. In each synthesis, one amino acid is added to LINK and the oligonucleotide sequence is extended by the corresponding code to give the following products:

CACATGGGGCCCTATTCTTAG–LINK–Gly
ACGGTAGGGCCCTATTCTTAG–LINK–Met

Step 3. The elongated products are pooled and again split into two parts for parallel synthesis, yielding

CACATGCACATGGGGCCCTATTCTTAG–LINK–Gly–Gly
CACATGACGGTAGGGCCCTATTCTTAG–LINK–Met–Gly
ACGGTACACATGGGGCCCTATTCTTAG–LINK–Gly–Met
ACGGTAACGGTAGGGCCCTATTCTTAG–LINK–Met–Met

Steps 4 and 5. Once more the products are pooled and divided into two aliquots for parallel synthesis. This results in an ensemble of eight tripeptide sequences, each encoded by a unique sequence of 18 nucleotides. The second PCR oligonucleotide is added to the ensemble of products to give

```
AGCTACTTCCCAAGGCACATGCACATGCACATGGGGCCCTATTCTTAG-LINK-Gly-Gly-Gly
AGCTACTTCCCAAGGCACATGCACATGACGGTAGGGCCCTATTCTTAG-LINK-Met-Gly-Gly
AGCTACTTCCCAAGGCACATGACGGTACACATGGGGCCCTATTCTTAG-LINK-Gly-Met-Gly
AGCTACTTCCCAAGGCACATGACGGTAACGGTAGGGCCCTATTCTTAG-LINK-Met-Met-Gly
AGCTACTTCCCAAGGACGGTACACATGCACATGGGGCCCTATTCTTAG-LINK-Gly-Gly-Met
AGCTACTTCCCAAGGACGGTACACATGACGGTAGGGCCCTATTCTTAG-LINK-Met-Gly-Met
AGCTACTTCCCAAGGACGGTAACGGTACACATGGGGCCCTATTCTTAG-LINK-Gly-Met-Met
AGCTACTTCCCAAGGACGGTAACGGTAACGGTAGGGCCCTATTCTTAG-LINK-Met-Met-Met
```

Implementation

Although natural amino acids are used in the example discussed above, the system is not limited to these, nor, for that matter, to peptides. The chemistry required for making encoded libraries is constrained only by the compatibility of the two alternating syntheses. Partly this involves the choice of the protecting groups, and the methods used to deprotect one chain while the other remains blocked. And, of course, each product needs to survive through the synthesis of the other. One can imagine many different ways of joining the chemical entities together, and one could even use mixed syntheses, provided that the rules of mutual compatibility are obeyed.

We have recently, in principle, solved the synthetic procedures for peptides (K. Janda, S. Ramcharitar, S.B., and R.A.L., unpublished results). Even within this field there is a choice of alphabets that extends well beyond the 20 natural α-amino acids. The only requirement is that we be able to make an amide bond. Thus, the amino and carboxylic groups can be located on a wide variety of compounds so that we can make libraries with many different backbone structures. We can also combine different backbones, if we define alphabets where, for example, both the number of carbon atoms and their configurations in the backbone are varied. New amino acids can be easily invented with unusual heterocyclic rings, such as thiazole-alanine or purine-alanine. These rings are components of natural effector molecules and often provide core chemical functions for important drugs. Libraries made with such alphabets will allow us to explore the combinatorial association of known effector chemical functions.

It is also useful to consider how large the combinatorial library should be. The PCR provides a very sensitive detection method, allowing even a few molecules to be seen. However, we need to have some reasonable concentration of each of the species present to cross the binding threshold of the acceptor molecule being assayed. If, for example, we set this as 1 μM and want 1 ml of the library, then we need to make at least 1 nmol of each of the species. Libraries with complexities of up to 10^4, giving us a total amount of 10 μmol of product, would seem reasonable. Because of this reciprocal relationship, more complex libraries could be made if the binding threshold is lowered.

Discussion

Traditional chemical synthesis proceeds by careful design, sequentially linking atoms or groups of atoms to a growing core structure. The process has the advantage that the product of each step can be analyzed, thereby allowing continuous evaluation of the effectiveness of a given strategy. Indeed, the analyzed results of these individual steps ultimately become the corpus of synthetic organic chemistry. A major technical revolution occurred with the advent of solid-state methods for the synthesis of polymeric molecules (10). Here, since a limited number of suitably protected oligomeric units are added via a common covalent bond, the results of the individual transformations can be predicted, and, to first approximation, it is necessary to analyze only the final product. In addition, the relationship of the monomeric units to each other and the extent of conformational space that is occupied can be estimated. Our method permits the study of the efficacy of combinatorial associations of diverse chemical units without the necessity of either synthesizing them one at a time or knowing their interactions in advance. It also allows easy identification of the most effective molecules through a common method of nucleic acid sequencing. Once the chemical polymers are decoded, more precise questions about critical interactions and conformations can be asked by reversion to classical chemical methods. Further, we expect that many receptors will interact with sets of related but not identical chemical entities such that major clues as to critical interactions can be deduced from the shared features of the sets.

Our method also provides a method of amplification, again by exploiting a common procedure of nucleic acid hybridization. In any screening procedure where large libraries of compounds or effector molecules are being studied, the absolute number of different nonspecific interactions may be large, but the specific ligand or effector is represented many more times than any individual background molecule. In such a situation the signal-to-noise ratio rapidly increases after repeated cycles of amplification and selection, and the specific molecule becomes highly enriched after only a few iterations. For both identification and selection, our method exploits the power of genetic systems. By coupling genetics and the versatility of organic chemical synthesis we have extended the range of analysis to chemicals that are not themselves part of biological systems.

We thank Kim Janda, Bernie Giluia, and Jerry Joyce for helpful comments on the manuscript.

1. Fodor, S. P. A., Read, J. L., Firrung, M. C., Stryer, L., Tsai, Lu, A. & Solas, D. (1991) *Science* 251, 767–773.
2. Geysen, H. M., Meloen, R. H. & Barteling, S. J. (1984) *Proc. Natl. Acad. Sci. USA* 81, 3998–4002.
3. Lam, K. S., Salmon, S. E., Hersh, E. M., Hruby, V. J., Kazmie[illegible] M. & Knapp, R. J. (1991) *Nature (London)* 354, 82–84.
4. Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C. T. & Cuervo, J. H. (1991) *Nature (London)* 354, 84–86.
5. Scott, J. K. & Smith, G. P. (1990) *Science* 249, 386–390.
6. Cwirla, S. E., Peters, E. A., Barrett, R. W. & Dower, W. J. (1990) *Proc. Natl. Acad. Sci. USA* 87, 6378–6382.
7. Tuerk, C. & Gold, L. (1990) *Science* 24, 505–510.
8. Ellington, A. D. & Szostak, J. W. (1990) *Nature (London)* 346, 818–822.
9. Crick, F. H. C., Griffith, J. S. & Orgel, L. E. (1957) *Proc. Natl. Acad. Sci. USA* 43, 416–421.
10. Merrifield, B. (1984) *Les Prix Nobel* (Almqvist & Wiksell International, Stockholm), pp. 127–153.